Australian Government
Department of Defence
Defence Science and
Technology Organisation

An Approach to Information Management for AIR7000 with Metadata and Ontologies

Csaba Veres and Simon Ng

Joint Operations Division

Defence Science and Technology Organisation

DSTO–TR–2289

ABSTRACT

This paper discusses the concept 'metadata', and shows its importance in information collection and dissemination activities. We also show that the information management components of maritime patrol and response mandate the effective use of metadata. We then propose an approach based on Semantic Technologies including the Resource Description Framework (RDF) and Upper Ontologies, for the implementation of metadata based dissemination services for AIR 7000. A preliminary architecture is proposed. While the architecture is not yet operational, it highlights the challenges that need to be overcome in any solution to the information management tasks of AIR 7000, and provides a possible form for the solution.

Published by

DSTO Defence Science and Technology Organisation
Department of Defence, Canberra 2600, ACT

Telephone: (02) 6256 6100
Facsimile: (03) 6256 6103

An Approach to Information Management for AIR7000 with Metadata and Ontologies

Executive Summary

The AIR 7000 project was set up to procure military systems to fill the capability gap left by the planned retirement of the AP-3C Orion aircraft, and to further enhance the maritime patrol and response capability of Defence. As part of the procurement, AIR 7000 will deliver both manned and unmanned aerial vehicles and associated infrastructure to support the tasking, collection, analysis and distribution of ISR data.

Joint Operations Division (JOD) is undertaking the Information Management Study (WBS 3.2.3) to support the science and technology research requirements for AIR 7000 as articulated in the AIR 7000 Science and Technology (S&T) Plan. The information management study is expressly concerned with examining the metadata requirements for the likely AIR 7000 information management approach outlined in the AIR 7000 Collection and Dissemination Architecture Study [Ng. *et al.* 2007].

The ability to find, access, edit and manage data in the Defence environment requires, at the very least, a means of identifying where and what data is available, and who has permission to see it. Metadata (and the capacity to annotate information, and retrieve information based on the confluence between tags and search requirements) is a potential enabler for all of these processes. Indeed, a number of AIR 7000 requirements directly call for metadata based tagging and retrieval.

However, simply mandating specific metadata schemas can result in interoperability problems. For example, many standards within the ADO mandate the use of XML for metadata markup, but generating time consuming XML annotations that conform to a specific prescribed schema is not always possible for every potentially interesting information object. Such objects might be marked up using improvised methods which are difficult to reconcile with the standards. To solve such problems, we propose an architecture in which different metadata schemes can inter operate. By using RDF (Resource Description Framework) as a common data model and Ontologies to provide common interpretation, fragments of metadata in various different base formats can be combined such that they contribute to, or combine with, extant metadata standards.

The conclusions of the report are listed on the following page:

- metadata implementation is important to support effective data management, dissemination and analysis within the AIR 7000 system;
- metadata implementation will only bridge the gap between users if existing and future metadata standards can be meaningfully related or mapped to one another;
- RDF is the recommended data model for storing data that can be combined in meaningful ways;
- Ontologies are a powerful method for mutually agreed interpretations;
- domain general "upper ontologies" provide one means for achieving mapping between metadata that is specific to a domain or community of interest;
- SUMO (Suggested Upper Merged Ontology) presents a viable upper ontology that could be used as the umbrella ontology for AIR 7000 and the wider Defence community.

Given these conclusions, the following recommendations are made:

- Work should be undertaken to identify appropriate domain metadata (current and future) for describing the various products (reports, plans, ISR data), processes and entities;
- AIR 7000 should commission a study to investigate how the relevant metadata standards should be made interoperable;
- RDF is the recommended data model for interoperating different standards;
- Ontologies are a recommended technology for achieving shared interpretation;
- SUMO should be considered as an upper ontology to support metadata interoperability;
- Future discovery services should adopt terms and processes consistent with the ontology adopted to support information management;
- An analysis of the practical risks associated with the use of metadata as a discovery enabler should be thoroughly examined.

We are currently working on methods to create ontologies from legacy data, and the use of the SUMO upper ontology for inter-relating the data sources. These general problems are being actively researched in the wider Information Systems community, but we are focusing on the problems specific to the defence domain. While we don't have complete working applications, we have prototyped components required for the architecture. Our research exposes some of the problems that need to be addressed in a deployed solution, and suggests some approaches to tackle the problems.

Authors

Csaba Veres
Joint Operations Division

Csaba finished his undergraduate degree at Monash University in 1987, after which he traveled to Tucson, Arizona, to obtain a Ph.D. in Cognitive Science. He returned to Australia in 1997, where he worked for DSTO in Adelaide (ITD). At the beginning of 2000 he left DSTO to work as a lecturer at Melbourne University in the Department of Information Systems. He then spent four years at the Norwegian University of Science and Technology in Trondheim, where he studied Ontologies and other Semantic Web technologies. He returned to Australia at the end of 2006 to join JOD, tackling the problem of information integration in Defence.

Simon Ng
Joint Operations Division

Simon Ng holds Bachelor of Science and Bachelor of Engineering degrees from Monash University. In 1998, he completed his doctorate of Philosophy in Materials Engineering. He undertook a post-doctoral fellowship for three years at the Commonwealth Science and Industry Research Organisation, where he developed methods for measuring the dielectric properties of cementitious systems. In 2001, Simon joined the Defence Science and Technology Organisation, supporting military experimentation and concept development as part of the Joint Experiment programme. Currently, he works for Joint Operations Division, where he is analysing information integration requirements for the AIR 7000 future maritime patrol and response project.

Contents

Figures

Tables

Glossary

ADO Australian Defence Organisation

ADO_DM_MDP Australian Defence Organisation Data Management MetaData Profile

DCMI Dublin Core Metadata Initiative

DDMS DoD Discovery Metadata Specification

DIE Defence Information Environment

DIGO Defence Imagery and Geospatial Organisation

DMO Defence Materiel Organisation

DOLCE Descriptive Ontology for Linguistic and Cognitive Engineering

EW Electronic Warfare

EXIF Exchangeable Image File Format

GPS Global Positioning Satellite

HP Hewlett Packard

HTML Hypertext Markup Language

IBM International Business Machines

IEEE Institute of Electrical and Electronics Engineers

IMINT Image Intelligence

ISO International Standards Organisation

ISR Intelligence Surveillance and Reconnaissance

JPEG Joint Photographic Experts Group

MIT Massachusetts Institute of Technology

MPEG Moving Picture Experts Group

NATO North Atlantic Treaty Organization

OWL Web Ontology Language

RDF Resource Description Framework

SIGINT Signals Intelligence

SUMO Suggested Upper Merged Ontology

UCO Upper Cyc Ontology

URI Uniform Resource Identifier

VDO Visual Descriptor Ontology

W3C The World Wide Web Consortium

XML Extensible Markup Language

XSLT Extensible Stylesheet Language Transformation

1 Introduction: Metadata in AIR 7000

The AIR 7000 project was set up to fill the capability gap left by the planned retirement of the AP-3C Orion aircraft, and to further enhance the maritime patrol and response capability of Defence. As part of the procurement, AIR 7000 will deliver both manned and unmanned aerial vehicles and associated infrastructure to support the tasking, collection, analysis and distribution of ISR data.

Joint Operations Division (JOD) is undertaking the Information Management Study to support the science and technology research requirements for AIR 7000 as articulated in the AIR 7000 Science and Technology (S&T) Plan[1]. The Information Management Study examines the metadata requirements for the likely AIR 7000 Information Management approach outlined in the AIR 7000 Collection and Dissemination Architecture Study (DSTO-CR-2007-0356) [Ng. *et al.* 2007].

A critical component of information management is enabling users to find, access, edit, and manage data from a number of different sources. This requires a means of identifying where and what data is available, who produced it, and who has permission to see it. Information of this sort is referred to as *metadata*, and is an enabler for required processes in the AIR 7000 information management system. Metadata is already a critical component of interoperability frameworks within Government. For example, the Australian Government Technical Interoperability Framework[2] includes a large number of metadata standards to enable interoperability. More generally, metadata based annotation and retrieval are an essential component of the digital knowledge workplace [Chase *et al.* 2006, Mack *et al.* 2001].

We note that there are also methods of finding and retrieving information based on more straightforward search and indexing approaches [Manning *et al.* 2007, Robertson 1994]. But these do not eliminate the need for metadata based approaches, since metadata is still relevant for access control, and other aspects of information management. In fact, even keyword indexes for a body of documents is a form of metadata (see for example table 2). In this report we only consider the role of metadata, and do not discuss general search techniques.

It is the aim of this study to consider the important aspects of implementing a metadata based approach to information management for AIR 7000. It will:

- introduce the concept of metadata in more detail;
- examine the AIR 7000 requirements in the context of the metadata discussion;
- introduce ontologies as an extension of metadata to enhance interoperability;
- highlight where and how metadata will form an important element in supporting the range of requirements for the information management system in AIR 7000;
- propose a preliminary solution for incorporating metadata practice into the Information Management model for AIR 7000.

[1] Australian Department of Defence 2007, Project AIR 7000 Science and Technology Plan (version 2.0), Canberra

[2] available at `http://www.agimo.gov.au/publications/2005/04/agtifv2`

2 An Introduction to Metadata

The term metadata means, literally, 'data about data'. This definition is meant to be broad enough to cover a wide variety of data "...necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system" [Gilliland 2008].

Our working definition draws on [Stephens 2004] who defines metadata as "Structured, semi-structured, and unstructured data which describes the characteristics of a resource (external source) or asset (internal source). Metadata is about knowledge, which is the ability to turn information and data into effective action." It is seen as an important enabler where systems are intended to support information pull, but also for the purposes of improving collaboration and interaction across a distributed enterprise. The literal definition might be thought of as "the sum total of what one can say about any information object at any level of aggregation".

Metadata can be broadly construed as capturing information about the three following aspects of data [Gilliland 2008]:

- **Content** relates to what the object contains or is about, and is intrinsic to an information object.
- **Context** indicates the who, what, why, where, how aspects associated with the object's creation and is extrinsic to an information object
- **Structure** relates to the formal set of associations within or among individual information objects.

Metadata can be viewed as a form of description to identify and provide access points to information objects, but also for documenting the administration, accessioning, preservation, and use of collections. The specific format of metadata differs according to use and context. Metadata can appear in HTML 'metatags' to make a Web site easier to find; in header fields of digitised images to record information about the image, the imaging process, and image rights; or embedded in applications to track information objects. For all its different uses, metadata is critical to identify and describe an information object, to document its function and use, its relationship to other information objects, and how it should be managed. Table 1 lists a number of possible metadata types, and provides some examples. Table 2 (on page 6) shows attributes of the metadata itself, and various possible values of those attributes. For example, the *source* of the metadata is an attribute that can have the values *external* or *internal*, depending on whether the metadata was attached to the data object as part of its creation, or attached later by a third party. (Both tables adapted from [Gilliland 2008]).

Table 1: *Common types of metadata, their definitions and examples.*

Type	Definition	Examples
Administrative	Metadata used in managing and administering information resources	- Security level - Acquisition information rights and reproduction tracking - Location information - Version control and differentiation between similar information objects - Audit trails created by record keeping systems
Descriptive	Metadata used to describe or identify information resources	- Annotations by users (content, relevance, etc.) - Cataloging records - Finding aids - specialised indexes - Hyperlinked relationships between resources
Preservation	Metadata related to the preservation management of information resources	- Documentation of physical condition of resources - Documentation of actions taken to preserve physical and digital versions of resources, e.g., data refreshing and migration
Technical	Metadata related to how a system functions or metadata behave	- Hardware and software documentation - Digitisation information, e.g., formats, compression ratios, scaling routines - Tracking of system response times - Authentication and security data, e.g., encryption keys, passwords
Use	Metadata related to the level and type of use of information resources	- Who is it useful for? - Tasking - Exhibit records - Use and user tracking

Metadata consists of a set of data that can be accrued over time to refer to many aspects of information objects. Some of this can be gathered automatically, but much of it is manual, or at least requires some manual intervention. This can be a time consuming and expensive process. However, [Gilliland 2008] argues that the requirements of information management in the digital age mandate the use of metadata. From the perspective of AIR 7000, the desire to support federated workflow and user-driven analysis and exploitation is enabled through a comprehensive metadata schema.

Some capabilities provided by metadata are discussed below:

- *Increased accessibility*
  Rich, consistent metadata can facilitate search and discovery and can also control access and editing privileges. In addition, resources from different sources can be spontaneously combined into virtual collections if metadata is used consistently across the various sites (that is, if a federated metadata catalog and an agreed or compatible ontology is implemented);

- *Retention of context*
  Repositories maintain collections of objects that often have complex interrelationships among each other, as well as associations with people, places and events. Metadata plays a key role in documenting and maintaining those relationships even in the event that the information objects are separated from their context. Similarly, metadata can indicate the structural and procedural integrity, and degree of completeness of information objects. As an example of dealing with original images and their copies, suppose one wanted to find an image of a Picasso painting from 1937 (from [Gilliland 2008]). The existence of metadata such as CREATOR = Picasso, DATE = 1937 is immediately helpful, but we must remember that the digital image is itself an information object with metadata like CREATOR = Scan-U-Like Imaging Labs Inc., DATE = 2000-02-29. So the context of the digital image must be kept separate from the original painting on which the image is based, to make sure that the original creator of the image depicted in the information object is findable;

- *Multi-versioning*
  The existence of information and cultural objects in digital form has heightened interest in the ability to create multiple and variant versions of those objects. Metadata is needed to link the multiple versions and capture what is the same and what is different about each version. The metadata must also be able to distinguish what is qualitatively different between variant digitised versions and the hard copy original or parent object;

- *Legal issues*
  Metadata allows repositories to track the many layers of rights and reproduction information that exist for information objects and their multiple versions. Metadata also documents other legal or donor requirements that have been imposed on objects - for example, security concerns or proprietary interests.

Table 2: *Attributes of metadata and possible values of those attributes.*

Attribute	Characteristics	Examples
Source of metadata	Internal metadata generated by the creating agent for an information object at the time when it is first created or digitised	- File names and header information - Directory structures - File format and compression scheme
	External metadata relating to an information object that is created later, often by someone other than the original creator	- Registrarial and cataloging records - User tags - Rights and other legal information
Method of metadata creation	Automatic metadata generated by a computer	- Keyword indexes - Source location information - User transaction logs
	Manual metadata created by humans	- Descriptive surrogates such as catalog records and Dublin Core metadata
Nature of metadata	Lay metadata created by persons who are neither subject nor information specialists, often the original creator of the information object	- Metatags created for a personal Web page - Personal filing systems
	Expert metadata created by either subject or information specialists, often not the original creator of the information object	- Specialised subject headings - Security rating - Archival finding aids
Status	Static metadata that never change once they have been created	- Title, provenance, and date of creation of an information resource
	Dynamic metadata that may change with use or manipulation of an information object	- Directory structure - User transaction logs - Image resolution
	Long-term metadata necessary to ensure that the information object continues to be accessible and usable	- Technical format and processing information - Rights information - Preservation management documentation
	Short-term metadata, mainly of a transactional nature	- Access times
Structure	Structured metadata that conform to a predictable standardised or unstandardised structure	- XML Schema - DDMS - local database formats
	Unstructured metadata that do not conform to a predictable structure	- Unstructured note fields and annotations
Semantics	Controlled metadata that conform to a standardised vocabulary or authority form	- DDMS - Dublin Core
	Uncontrolled metadata that do not conform to any standardised vocabulary or authority form	- Free-text notes - HTML metatags
Level	Collection metadata relating to collections of information objects	- Collection-level record, e.g., task based - Specialised index
	Item metadata relating to individual information objects, often contained within collections	- Transcribed image captions and dates - Format information

3 Requirements Drivers in AIR 7000

The requirements discussed in this section are derived from [Ng. *et al.* 2005]. Table 3 indicates which AIR 7000 requirements mandate or, at the very least, suggest a need for metadata.

Requirements that mandate metadata are self-explanatory, because they specifically ask for "metadata" (e.g. A014) or they specifically require a function that depends on the availability of metadata. For example requirement S002 is about the origin and versioning of data, which is contained in the metadata. These requirements could not be met without implementing a metadata schema.

Requirements that suggest a need for metadata could potentially be achieved by other means, but metadata either offers a simpler implementation or adds features that enhance the original requirement. For example, "D007: The system shall support effective dissemination of data to/from non-Defence organisations" could be implemented at some level without metadata. However the use of metadata is required for a flexible system of dissemination with access controls, version management, traceability, and so on. A common set of terms and interoperable data standards results in a useful interoperability framework.

Table 3: *AIR 7000 requirements that either mandate or suggest the adoption of a metadata approach to support information management.*

ReqID	Requirements. *The system shall*:	Mandate	Suggest
A003	allow analysis of ISR to be conducted across the federated information environment	*	
A012	automatically generate standardised metadata at the point of data creation (where possible) using sources of available data (such as GPS information, etc)	*	
A014	facilitate effective and efficient manual tagging of data (including streaming video) in near real time using standardised metadata/tagging schema	*	
C007	allow an information user to assess the degree to which information can be trusted as being accurate.	*	
C016	support a common/translatable lexicon, taxonomy and indexing to enable effective interoperability within Defence and Australian non-Defence government	*	
D005	provide services for automated notification, retrieval and dissemination of data, using subscription, profile push and brokering services as appropriate.	*	
R006	provide users with search, discovery and subscription services and infrastructure to allow information pull from the federated data environment (including Defence, non-Defence, allied and internet sources)	*	
R018	support the effective and efficient retrieval of non-current (such as (including all communication data) across the federated environment in near real time	*	
S002	support traceability and visibility of data (including raw data, product and non-intelligence data, such as RFIs) origin, processing and versioning across the federated data environment in accordance with rules of access.	*	
S011	enforce the archiving of data in accordance with legal and operational requirements	*	
S016	be interoperable with current and planned future data stores and applications (including databases, support systems, management systems, machine)	*	
A001	provide the ability for users to fuse and/or correlate (automatically & manually) data across multiple federated sources		*
A015	support situational awareness tools with layered data to provide a means for drilling into underlying information (including location, track information and status)		*
A016	allow users to ‘mash’ multiple sources into a personalised situational awareness display		*
D002	allow dissemination of preprocessed data and intelligence product (including sensor data (IMINT, SIGINT, EW, track information), briefings and other fused data)		*
D007	support effective dissemination of data to/from non-Defence organisations		*
D009	store, retrieve and disseminate intelligence product in appropriate/required formats/standards.		*
D012	disseminate information in accordance with access rules for users based on measures of risk and rules of access (including security classification and caveats, needs-to-know, consumer role, etc).		*
S010	have effective manual and automated data management capability, including the ability to automatically index and categorise, accept, upload and store intelligence data from relevant federated sources.		*
S013	allow contradictory data to be evaluated and shall manage the problem of versioning control, configuration management, editing permissions, circular reporting duplicate analysis, etc. in order to ensure the traceability, reliability and quality of data		*

3.1 An assessment against requirements

The need for metadata is clearly indicated by the requirements in Table 3 where current stakeholders have already identified the requirement for metadata as a need to the performance of certain tasks. In addition, there are many requirements in which stakeholders have not identified metadata as a solution, but have expressed the need to perform tasks which typically involve metadata. For example, S011 requires that data be archived according to legal and operational requirements. But archiving normally involves metadata. At the simplest level we assume at least a creation date, or location, or author, or some characteristic by which a number of information resources can be grouped for storage or retrieval. As the sophistication of these requirements increases, so does the need for metadata. As another example, requirement A016 states that users should be able to 'mash' multiple sources into a personalised display. While it is possible to imagine a simple version of this capability without metadata, a sophisticated application would allow users to express their personalised needs in a comprehensive and rich framework which would then be able to link to appropriate data. It is difficult to imagine how this could be achieved without metadata.

In summary the requirements dictate the need for a metadata scheme that:

- is effective at allowing end-users and analysts to have relevant data brought to them based on the confluence of metadata and their own preferences/needs as users - which also minimises handling;
- will support the pull of information using automated and sophisticated approaches (such as subscription and brokering);
- will help to manage storage and archiving of the vast amount of data within the federated information environment;
- will enable tracing of data source, versions, assurance;
- will allow enrichment of data by giving users the capacity to value-add to existing metadata;
- will support fusion, enhanced situational awareness tools and collaborative spaces; and,
- is more demanding on processes (both manual and automatic) and on a cultural shift to its rigorous use.

4 Recommended Metadata Schema for the AIR 7000 Information Environment

A large number of metadata schema currently exist that might be useful for supporting information management processes in Defence and more specifically within the user community of relevance to the future AIR 7000 capability. While this is positive in terms of finding existing solutions, it also creates some problems. First, the kinds of concepts included in each schema, and their level of generality, is typically determined by the intended application. While there are a large number of possible metadata schemas to chose from, there is rarely a 'perfect' one to use for a new application. Secondly, the different schemas are often expressed in different syntactic formalisms, often as plain text or as XML Schema specifications. To combine these schemas, it is first necessary to translate them to a common format. Thirdly, systems and processes may be designed to cope with current or planned schema, and to adopt a new schema implies possible associated transition costs and effort.

4.1 The Metadata Format

The W3C-recommended data model for semantic web applications is the Resource Description Framework (RDF) [3]. RDF is a deceptively simple data model whose base structure is a triple, where the elements in the triple are often thought of in terms of a linguistic analogy: subject - predicate - object. Each element in the triple must be a resource identified by a Uniform Resource Identifier (URI), except the object, which could also be a literal (such as a string or a number). In essence, each triple makes a simple statement about a resource, of the form AIRPLANE *type* GLOBAL HAWK, AIRPLANE *type* SUPER HORNET. But because each element in a given statement is also a resource, we can easily add further statements like for example GLOBAL HAWK *has capability* RADAR, and so on. The data model is therefore a flexible and extensible one, which allows complex networks of facts to be accumulated, and inferences drawn from these facts. For example the set of statements just given allows us to ask for all the capabilities of all our airplanes, or perhaps more usefully, we could ask about which airplanes possess a given capability. The addition of *namespaces* allow us to be precise about which resources we might be making statements about. In the current example the English word AIRPLANE is not sufficient for a unique reference, so we qualify it with a namespace like `http://defence.gov.au#airplane`, which gives it a unique name to fix reference. Figure 1 is an example RDF graph from the W3C RDF Primer document[4], which shows a fully namespace qualified set of statements that supply information about the resource identified by `http://www.w3.org/People/EM/contact#me`.

The complete set of RDF triples tells us several things about the resource in the example, including that its full name is Eric Miller. But we also know that the current understanding of *full name* is precisely `http://www.w3.org/2000/10/swap/pim/contact#fullName`. This is a property defined in the document found at `http://www.w3.org/`

[3] `http://www.w3.org/RDF/`

[4] `http://www.w3.org/TR/rdf-primer/`

`2000/10/swap/pim/contact`, which defines this precise meaning of fullName. Such a definition could include whether or not a middle name should be given, if the surname should come first, and so on. The fullName property is therefore unambiguously defined in figure 1, and will not be confused with a different definition that might be given elsewhere.

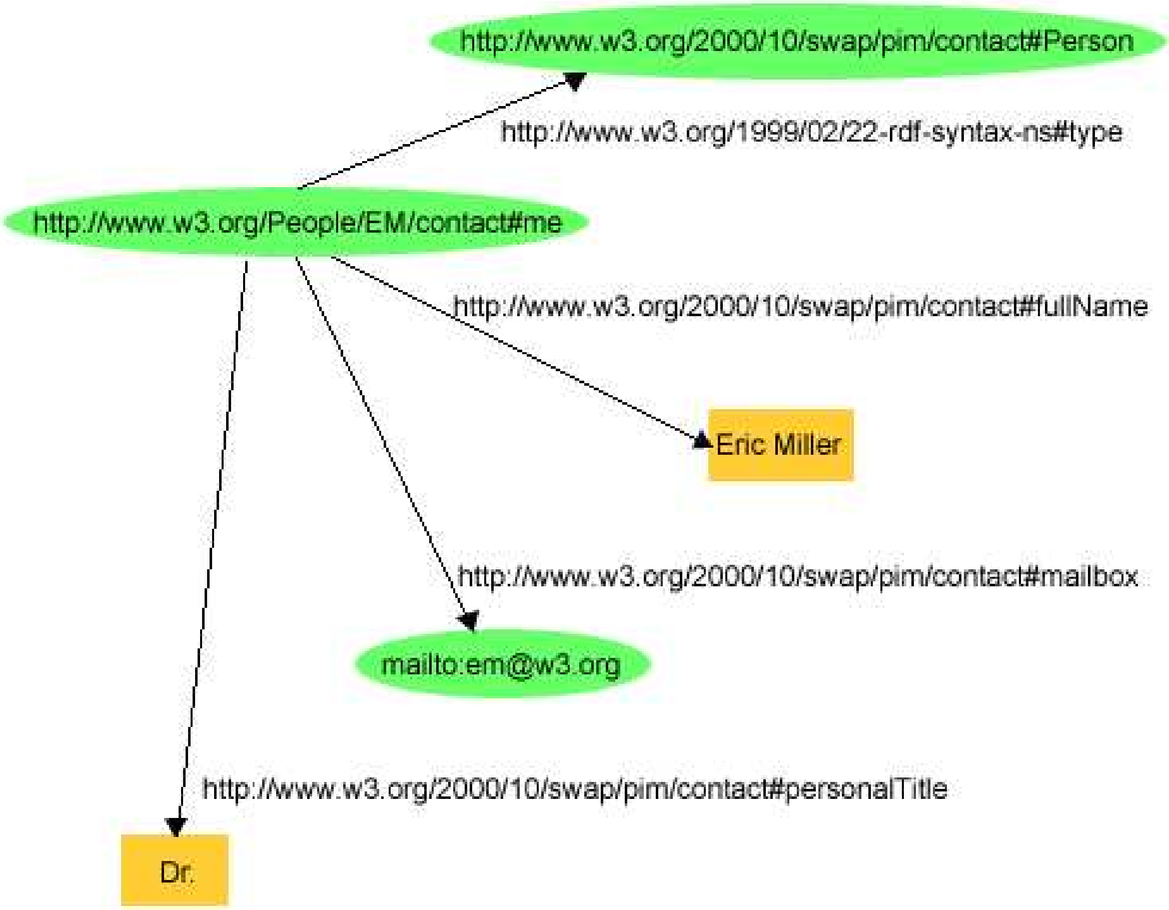

http://www.w3.org/2000/10/swap/pim/contact#Person
http://www.w3.org/1999/02/22-rdf-syntax-ns#type
http://www.w3.org/People/EM/contact#me
http://www.w3.org/2000/10/swap/pim/contact#fullName
Eric Miller
http://www.w3.org/2000/10/swap/pim/contact#mailbox
mailto:em@w3.org
http://www.w3.org/2000/10/swap/pim/contact#personalTitle
Dr.

Figure 1: *A RDF graph of triples 'http://www.w3.org/People/EM/contact/#me'*

The RDF approach has several advantages over standard relational or flat file structures:

- Combining triples leads to a potentially vast 'semantic network' where units of data can link in complex ways to various resources;
- Information about a given resource can be stored anywhere on the federated network; and,
- New kinds of information about a resource can be added dynamically without the need to alter an underlying data model.

Existing schema of various sorts should initially be mapped as RDF in order to realise the potential of RDF with existing schema. As we will see in subsequent sections, many existing standards require that the metadata is expressed as XML to conform to some XML Schema definition. This makes the accurate translation of XML into RDF a task of primary importance[5]. The viability of undertaking this mapping process for information sources and data types in Defence should be investigated further.

RDF is currently not a mainstream technology in industrial applications. Its origins are from the Semantic Web initiative, which is an anticipated evolution of the Internet in which web sites will contain data that can be consumed by machines as well as humans. RDF fits into this initiative because it provides a flexible distributed data model in which any source from around the world can add information to data held at any other source on the Internet. While Semantic Web technologies are not yet mainstream, major technology companies (e.g. IBM, HP) and research establishments (e.g. Stanford University, MIT, Oxford) have invested a great deal of resources into Semantic Web related research activities, and several data integration companies have begun using the technologies[6].

The advantages to defence intelligence can be summarised in figure 2, which presents a simplified view of three RDF fragments known from data stores around the world. First, we know from a data store in Asia that there was a weapons shipment to Sudan on October 10 1998. We also know from a store in France from a world immigration database that Al-Amir visited Sudan on October 10 1998. But a data store in Washington identifies that Al-Amir is an alias for Osama bin Laden, so immediately we can infer that bin Laden was in Sudan on the day of a major weapons shipment, and therefore postulate that he was involved in this deal. Such a concatenation of data would not be possible in existing systems where data resides in closed, proprietary data formats that could not easily be shared and combined.

[5] To make this easier, there are a number of existing tools to map various metadata formats to RDF. MIT's SIMILE project maintains a list of 'RDFIzers' to translate various formats to RDF. These include JPEG, BibTEX, Email, Weather, Outlook, EXIF. It is also possible to map XML and XML schema documents to RDF using XSLT. This is particularly important because many government and defense metadata is in terms of XML schema.

[6] e.g. `http://www.insilicodiscovery.com/v2/index.php`, `http://www.talis.com/platform/index.shtml`

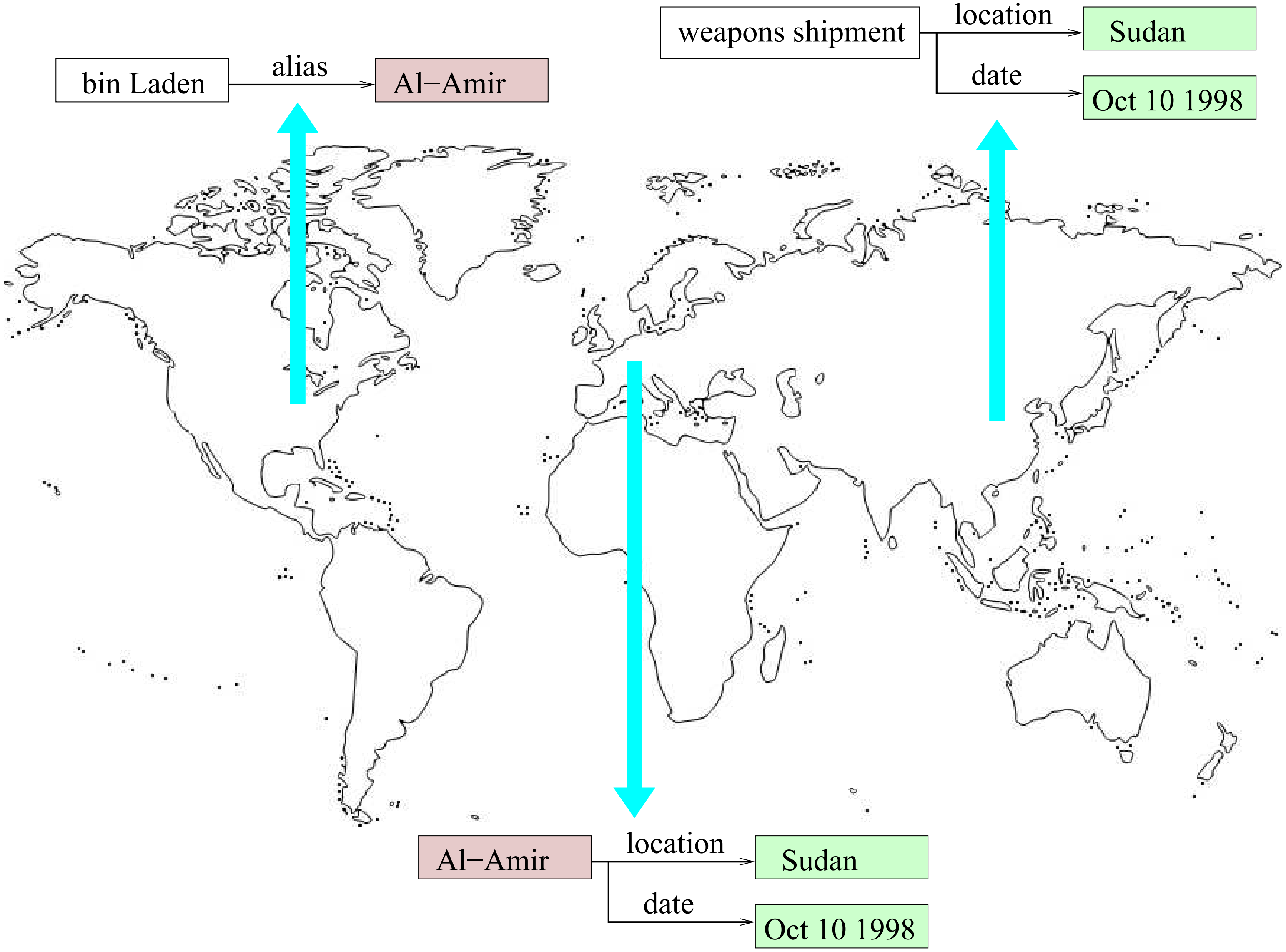


Figure 2: *RDF triples stored in data stores around the globe.*

5 Introduction to Ontology

Ontologies provide a way to fix the interpretation, or semantics of the assertions. For example, we saw in the previous section that the resource `http://defence.gov.au#airplane` is an unambiguous name for a resource we want to reference. However we cannot say anything about the relationship of this resource to, for example, `http://www.defense.gov#airplane`, which might be a term defined in the United States Department of Defense. While the English word for the air asset, which is used by the two organizations is the same, there is no way to formally know that the range of real world objects referred to by the two terms is the same. Ontologies typically make it possible to define concept equivalences, and other useful relationships.

The de facto standard for constructing ontologies is the W3C recommendation for the Web Ontology Language (OWL)[7]. OWL contains the basic elements owl:Class and owl:Property, which are used to build basic terminologies with classes of objects and properties associated with classes. But it also contains logical connectives like owl:sameAs, which can express equivalence between concepts as in the previous example, and owl:subClass, which specifies an inclusion relation between two concepts. From the definition of owl:subClass as a transitive property we can infer, for example, that if *FA/18* is a *subClass* of *fighter plane*, and *fighter plane* is a *subClass* of *plane* then *FA/18* is a *subClass* of *plane*. This is a very simple example, but the point is that ontologies can be used to build rich domain models that express logical relationships between concepts of interest. The reference and semantics of these domain models is fixed by the various mechanisms available in RDF and OWL.

5.1 Upper Ontologies

The flexibility of modeling with ontologies allows different stakeholders to model similar real world entities with different terminologies. This is a prevalent problem for metadata in general, where stakeholders often produce unique vocabularies for their specific needs. Defence (and the wider community involved in Defence operations) use different concepts and terms to describe similar sorts of objects, which hinders large scale interoperation between standards using different terminologies. Several interviewees in the AIR 7000 requirements exercise identified the need for the information system and the communities it services to have a shared lexicon or some method of translating between lexicons (e.g. requirement C016).

In this section we outline a proposed technology for inter operating vocabularies and data using a high level, domain independent upper ontology as an abstract, common top layer that ties together individual domain ontologies (that is, ontologies that are used to describe information within a community of interest). The IEEE P1600.1 Standard Upper Ontology Working Group defines the usefulness for upper ontologies " ... to support computer applications such as data interoperability, information search and retrieval, automated inferencing, and natural language processing."[8]

An upper ontology gives a domain independent, high level view of the sorts of con-

[7] `http://www.w3.org/TR/owl-features/`

[8] http://suo.ieee.org/SUO/scopeAndPurpose.html

cepts that exist in the world, such that these concepts then subsume all the concepts in the domain ontologies. It is obviously not possible to give a completely objective and impartial view of the world, so all proposed upper ontologies begin with some philosophical bias such as the distinction between tangible objects and substances from which they are made, the difference between objects and their properties, and so on. This raises the problem that different upper ontologies have different biases, and may be incompatible. The choice of a useful upper ontology early in the project is therefore critical. Given a set of high level categories, concepts from domain ontologies can be thought of as sub classes of the high level categories. This is useful for comparing entities from different domain ontologies, because if two domain concepts map onto the same upper ontology category, then we can infer that the two domain concepts are similar. Consider the following brief example involving metadata schemas used in defence:

Two important metadata standards (explored in more detail later) are the Dublin Core Metadata Initiative (DCMI) and the DoD Discovery Metadata Specification (DDMS). These define the dc:creator and the ddms:creator properties, respectively. These properties are subtly different in detail, but both can be mapped to the property sumo:authors in the upper ontology: Suggested Upper Merged Ontology (SUMO). Suppose a specific query comes in from a user familiar with Dublin Core, with a format such as 'find all documents with dc:creator Adam Smith'. Clearly this would miss potentially relevant documents that were created by Adam Smith under the DDMS schema (ddms:creator). But if the user was unaware of the distinction between DCMI and DDMS (or any other such schema) and only knew about the upper ontology SUMO, he or she would search for 'authors' of a given text. Figure 3 shows a fragment of the SUMO ontology that shows the relationship between a text and its author. If we wanted to find all humans who authored texts we could specify an appropriate query on the SUMO ontology, and since both dc:creator and ddms:creator were sub properties of sumo:authors, the query would return answers marked up by either metadata standard. In this example, the upper ontology serves as a user friendly interface to the rich metadata in the system.

A related benefit of using an upper ontology is to clarify the relationships between metadata standards in a given framework, as they apply to particular information objects. For example the Australian Government Technical Interoperability Framework includes a number of standards for image data, such as JPEG, MPEG-1, MPEG-2, and so on. An upper ontology would clarify the relationship of these standards to the information objects present in a system, and make explicit the details of the standards in such a way that overlapping information could be exploited in information retrieval.

MITRE corporation has considered the applicability of Upper Ontologies for military use [Semy *et al.* 2004]. They suggest a number of ontologies, but focus on arguably the three most well known of these: the Suggested Upper Merged Ontology (SUMO), Upper Cyc Ontology (UCO) and Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE). MITRE Corporation's review emphasises that upper ontologies are constructed with particular assumptions and theoretical approaches. They embody different views of the world, which might themselves be incompatible[9]. It is therefore critical to choose an appropriate upper model for an intended domain and set of required use cases.

[9] Ironically, two domain ontologies constructed on the basis of two different upper ontologies might turn out to be incompatible, even though upper ontologies are supposed to help with interoperability.

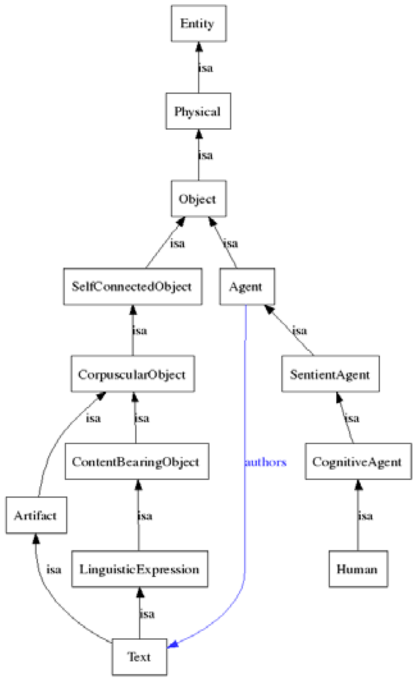


Figure 3: *Part of SUMO Ontology illustrating connection between author and text.*

The report lists four additional assessment criteria for evaluating candidate upper ontologies for military domains:

- *Licensing*
  It is argued that an open license is critical in a Government domain since open standards facilitate interoperability and information sharing across Government organisations as well as with coalition partners;

- *Structure*
  should allow extensibility and flexibility. A modular design is preferable since it facilitates reuse, extensibility, and community contribution;

- *Maturity*
  there should be a certain level of reliability and acceptance, to mitigate risk in operational environments as well as investment in development effort; and,

- *Miscellaneous*
  precision, security.

[Semy *et al.* 2004] are predisposed towards DOLCE as a foundational ontology for constructing domain models. The following sections briefly introduce some of the competing upper ontologies, and argues that SUMO is in fact a better choice for application to the AIR 7000 information management system, and to Defence in general.

5.1.1 Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE-Lite)

DOLCE [Masolo *et al.* 2007] is a theoretical project conducted as part of the EU sponsored WonderWeb project, and now hosted at Laboratory for Applied Ontology in Italy[10]. It is not a single upper level ontology, but serves as the topmost ontology for a library of foundational ontologies. Figure 4 illustrates a small number of the top level concepts in DOLCE.

The ontology embodies a large number of philosophical and logical considerations, though it is often claimed to be cognitively inspired. The two fundamental distinctions in the ontology contrast particulars with universals and endurants with perduants. The ontology is about particulars, or 'entities' as commonly known. The nodes in the ontology represent classes of such particulars. The DOLCE category structure is extremely complex and is defended only through complex philosophical justifications, which we find problematical from a usability standpoint, at least in the domain of human centred information retrieval.

5.1.2 The Suggested Upper Merged Ontology (SUMO)

SUMO is a suggested upper ontology that was constructed from a merge of existing publicly available schemas and synthesised under the guidance of the IEEE working group. It thus embodies consensus from contributors with a large number of theoretical orientations from the fields of engineering, philosophy, and information science [Niles and Pease 2001]. Figure 5 illustrates a few concepts from the top level of SUMO.

The topmost concept Entity subsumes Physical and Abstract. The former category includes everything that has a position in space/time, and the latter category includes everything else. Physical entities can be either Object or Process. Because SUMO was synthesised from the common features of a number of approaches, it seems relatively intuitive without technical explanations.

5.1.3 Upper Cyc Ontology

The UCO is abstracted away from the concepts in the Cyc project, which predates the current interest in ontologies. It was originally construed as a massive knowledge based system encompassing a vast amount of human 'common sense reasoning'. It is probably the largest project in existance for formalising human knowledge for the use of automated reasoning. Cyc itself is a complex knowledge base of 'common sense' knowledge, and the upper ontology is an attempt to introduce categorical structure to this knowledge. Figure 6 again provides an illustration of the complexity of the Cyc ontology model.

[10] `http://www.loa-cnr.it/DOLCE.html`

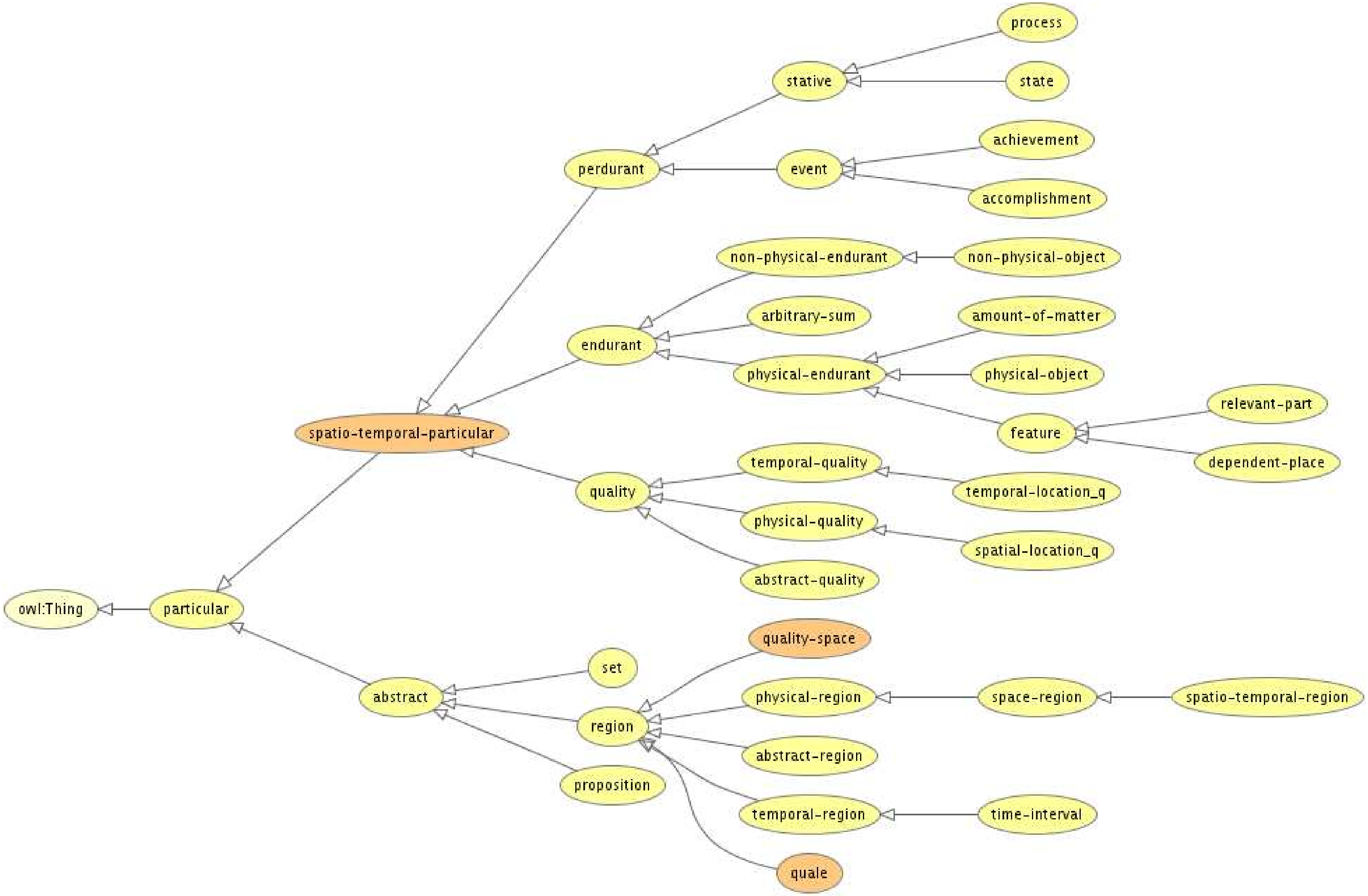


Figure 4: *Selection of DOLCE top level concepts.*

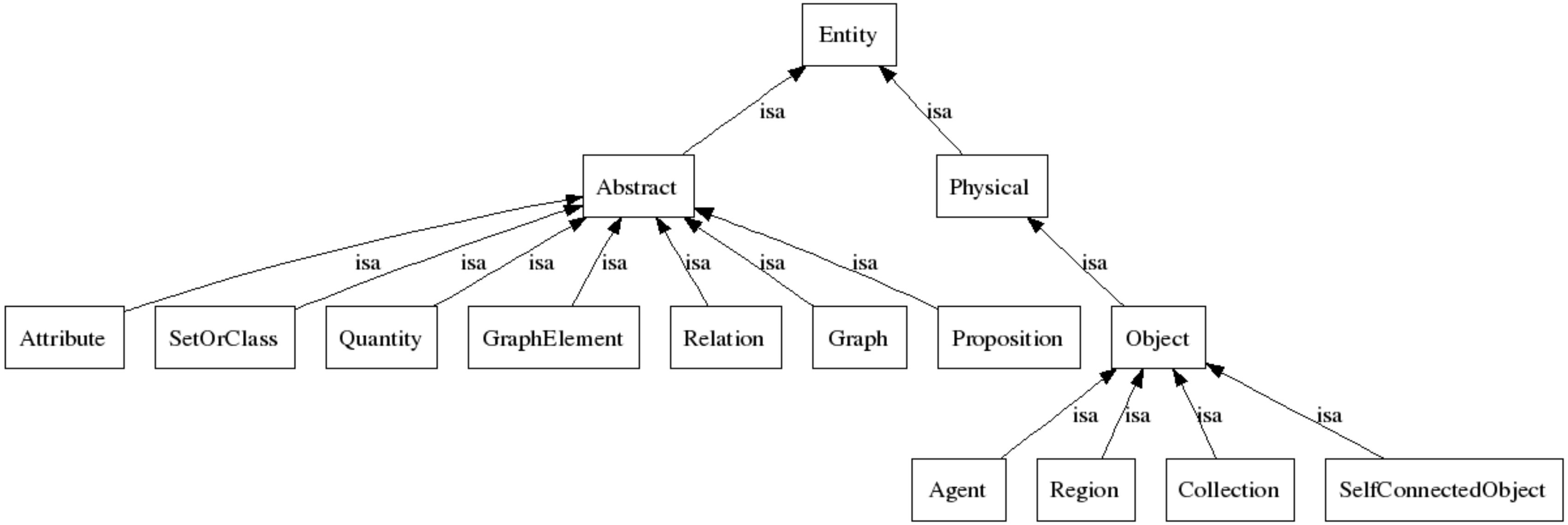


Figure 5: *Selection of SUMO top level concepts.*

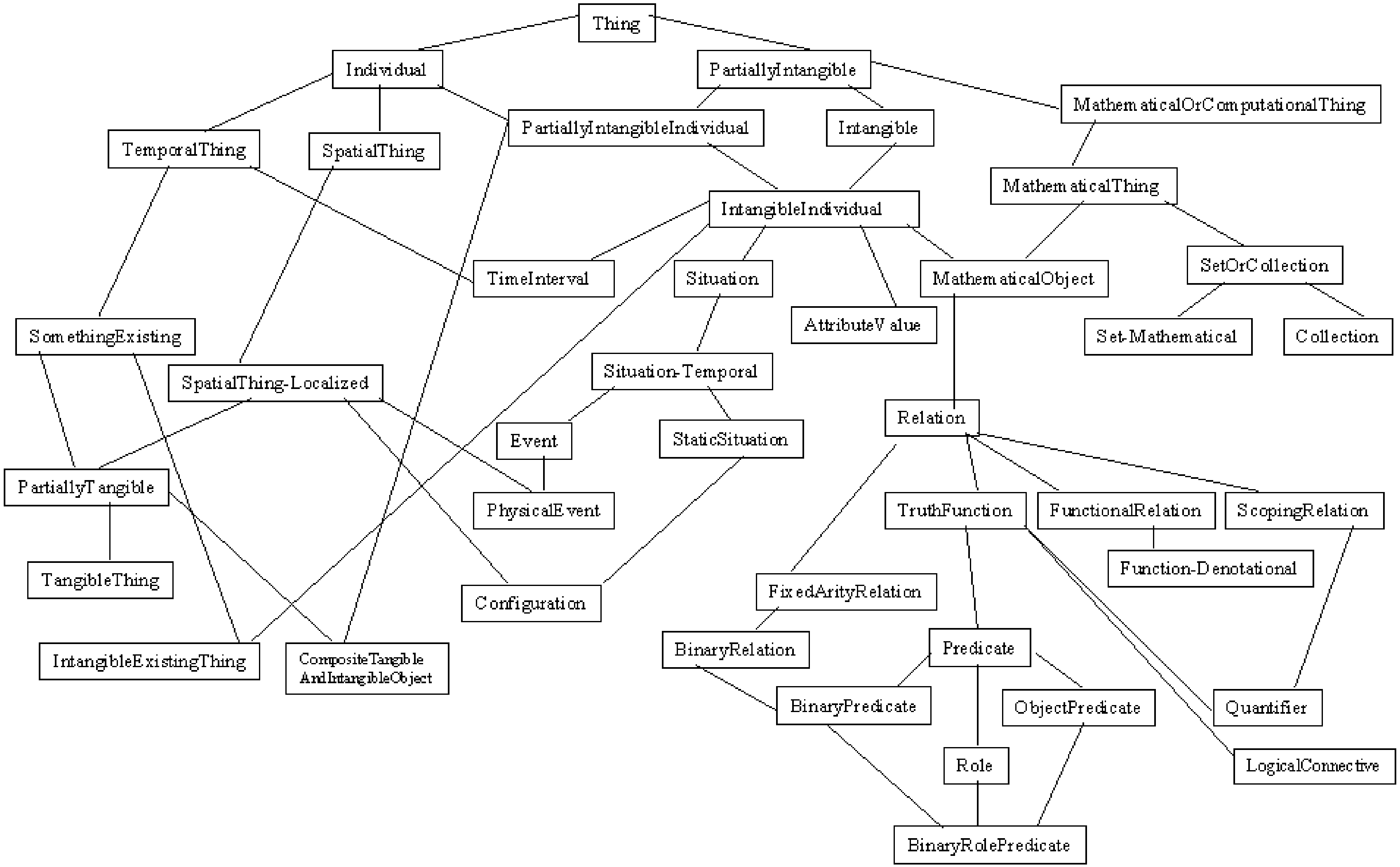


Figure 6: *Cyc upper level ontology.*

5.2 Summary

The choice of the most appropriate upper ontology is critical, since it impacts on a large number of subsequent decisions. In their report [Semy *et al.* 2004] conclude that DOLCE and SUMO are more or less on par according to their Defence specific evaluation criteria. However they are inclined toward DOLCE because it " ... is in general better informed by formal ontological analysis and formal semantics ... "

DOLCE is a complex framework that embodies formal theoretical principles in ontology engineering (most notably OntoClean)[11]. But [Sowa 2008] argues that an upper ontology with fewer semantic assumptions might be less problematic because it does not force the user into premature ontological commitments.

DOLCE is complex in part because it was designed to support automated negotiation, which differs in scope from the sorts of requirements we have in AIR 7000. On practical grounds we suggest that the assumptions embodied in DOLCE are less than transparent, and we disagree with the assessment in [Semy *et al.* 2004].

This report recommends that SUMO provides an intuitive upper model for Defence's domain ontologies, and we are currently undertaking work in which we are mapping important defence ontologies to SUMO.

[11] The philosophical assumptions built in to the ontology are also at times questionable. For example the fundamental distinction between endurant (endures for all time) and perduant (temporary) is not entirely intuitive, and is under vigorous philosophical debate [Varzi 2000]

6 An Overview of Relevant Domain Ontologies/Metadata

The number of domain specific metadata schemas is large, so our recommendations cover a small but highly visible subset. These are some of the most important of the available schemes that will be used for representing specific metadata within a domain or community of interest. Some of the metadata schemes are already expressed as ontologies or have been translated into RDF format. This is true for the majority of the publically used schemas, but less so for the defence specific ones. As a result, part of our on going work involves the creation of techniques for translating defence schemas into RDF for possible inclusion in Ontologies.

6.1 Australian Defence Organisation Data Management MetaData Profile (ADO_DM_MDP)

The Defence Imagery & Geospatial Organisation (DIGO) Standards Office maintains a set of policy documents that identify standards to be used in the management of geospatial information. The ADO_DM_MDP[12] is a specialisation of the ISO 19115:2003 Geographic Information – Metadata standard[13] which "... defines the schema required for describing geographic information and services" and "... provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data".

The ADO_DM_MDP defines a profile that is a specialisation of the ISO standard, designed especially for use by the Australian Defence Organisation. The DIGO Policies site[14] introduces the Metadata Profile as follows:

> "Interoperability in the Australian Defence Organisation (ADO) dictates the necessity for standardised practices in the collection of metadata. This document identifies the core metadata elements considered necessary to facilitate discovery, access and evaluation of data across the Defence Information Environment (DIE)."

In order to conform to the profile, metadata should be provided in XML format and conform to the XML Schema Definitions provided with the profile. The XML Schema definitions are relatively complex, with a large number of imports and dependencies between the individual modules. The useful translation of these standards to RDF is a non trivial task, but will have a number of benefits for the inherent interoperability of the data model. If the reference to the information object can be uniquely fixed, then different elements of the metadata definitions can be combined in a piecemeal fashion. For example, the *contact* information for the maintainer of the document can be specified in one location and detailed *spatial representation information* in another, but this will be invisible to

[12] http://intranet.defence.gov.au/digo/docs/ADO_DM_MDP_V1.04.pdf

[13] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=26020

[14] http://intranet.defence.gov.au/digo/stds_policy_docs.html

the combined metadata definition. Another consequence is that other documents that have the same *spatial representation information* can also be identified instantly, without complex searches through XML documents. This becomes even more useful when other metadata standards (for example, DDMS discussed below) define *spatial representation information* in the same way, because then searches will return documents annotated by either ADO_DM_MDP or DDMS (in this example). In addition, documents can be recovered on the basis of arbitrarily defined relationships between properties. For example, one could define *areas of interest* by enumerating a number of different instances of *spatial representation information*, and then retrieve documents based on *areas of interest*.

Migrating the metadata from a flat XML file to an RDF data source can make the same metadata available to a wider set of applications, and also allows semantically enriched searches to be performed over a federated source.

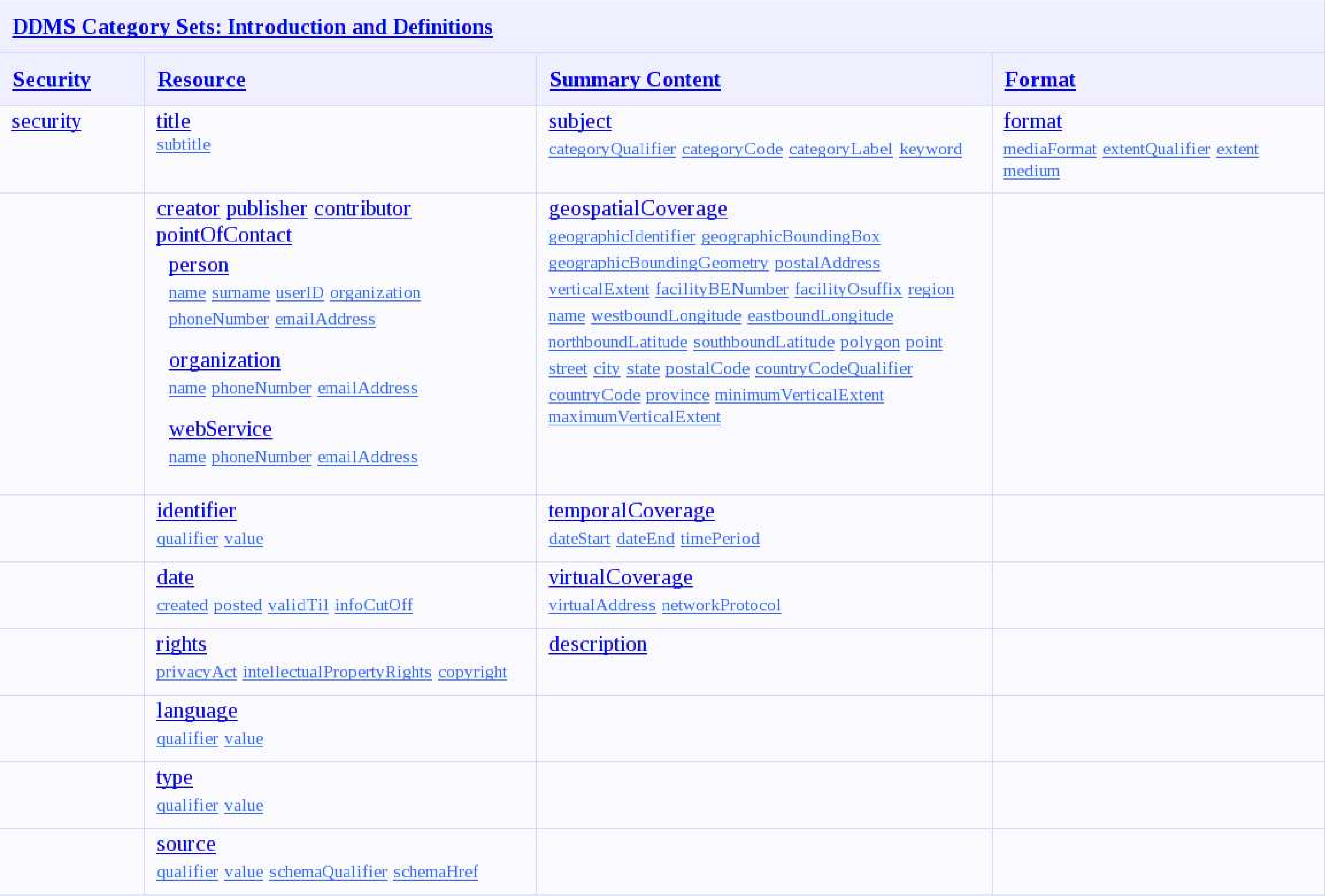
6.2 DDMS (US DoD)

The US Department of Defense Discovery Metadata Specification (DDMS)[15] defines discovery metadata elements for documents posted to community and organisational shared spaces. It is specified as a set of XML Schema.

'Discovery' is defined in the DDMS as the ability to locate data assets through a consistent and flexible search. Visibility, accessibility, and understandability are the high priority goals of the DoD Net-Centric Data Strategy. With the express purpose of supporting the visibility goal of the DoD Net-Centric Data Strategy, the DDMS specifies a set of information fields that are to be used to describe any data or service asset, i.e., resource, that is to be made discoverable to the Enterprise, and it serves as a reference for developers, architects, and engineers by laying a foundation for Discovery Services. The DDMS will be employed consistently across the Department's disciplines, domains and data formats.

The DDMS includes the Dublin Core Metadata elements (see below). For example, the creator element is defined using the DCMI core elements. Figure 7 shows the DDMS metadata elements that define document metadata to do with the creation, contents, security rating, and other administrative aspects of a document.

[15] `http://metadata.dod.mil/mdr/irs/DDMS/`



DDMS Category Sets: Introduction and Definitions

Security	Resource	Summary Content	Format
security	title subtitle	subject categoryQualifier categoryCode categoryLabel keyword	format mediaFormat extentQualifier extent medium
	creator publisher contributor pointOfContact person name surname userID organization phoneNumber emailAddress organization name phoneNumber emailAddress webService name phoneNumber emailAddress	geospatialCoverage geographicIdentifier geographicBoundingBox geographicBoundingGeometry postalAddress verticalExtent facilityBENumber facilityOsuffix region name westboundLongitude eastboundLongitude northboundLatitude southboundLatitude polygon point street city state postalCode countryCodeQualifier countryCode province minimumVerticalExtent maximumVerticalExtent	
	identifier qualifier value	temporalCoverage dateStart dateEnd timePeriod	
	date created posted validTil infoCutOff	virtualCoverage virtualAddress networkProtocol	
	rights privacyAct intellectualPropertyRights copyright	description	
	language qualifier value		
	type qualifier value		
	source qualifier value schemaQualifier schemaHref		

Figure 7: *The DDMS Specification, core elements.*

6.3 Dublin Core Metadata Initiative (DCMI)

The Dublin Core Metadata Initiative "provides simple standards to facilitate the finding, sharing and management of information"[16], and defines basic document identification metadata. It is perhaps the most mature and widely accepted Internet based metadata schema, and is included in many other schemas such as the DDMS. The specification involves a basic set of 15 elements in the Simple Dublin Core, and adds extensions and refinements in the Qualified Dublin Core. The 15 basic elements provide metadata sufficient to describe the origin and basic type of a document. They are:

- *contributor:* an entity responsible for making contributions to the resource;
- *coverage:* the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant;
- *creator:* an entity primarily responsible for making the resource;
- *date:* a point or period of time associated with an event in the lifecycle of the resource;
- *description:* an account of the resource;
- *format:* the file format, physical medium, or dimensions of the resource;
- *identifier:* an unambiguous reference to the resource within a given context;
- *language:* a language of the resource;
- *publisher:* an entity responsible for making the resource available;
- *relation:* a related resource;
- *rights:* information about rights held in and over the resource;
- *source:* the resource from which the described resource is derived;
- *subject:* the topic of the resource;
- *title:* a name given to the resource; and,
- *type:* the nature or genre of the resource.

Since the DCMI defines metadata at a very general level, it has some basic extension sets to provide more detailed metadata information. Figure 8 shows extensions to some concepts in the basic DCMI element set. For example, the node *TypeScheme* extends *type* and provides a number of sub elements like *text* and *image*.

[16] http://dublincore.org/about/

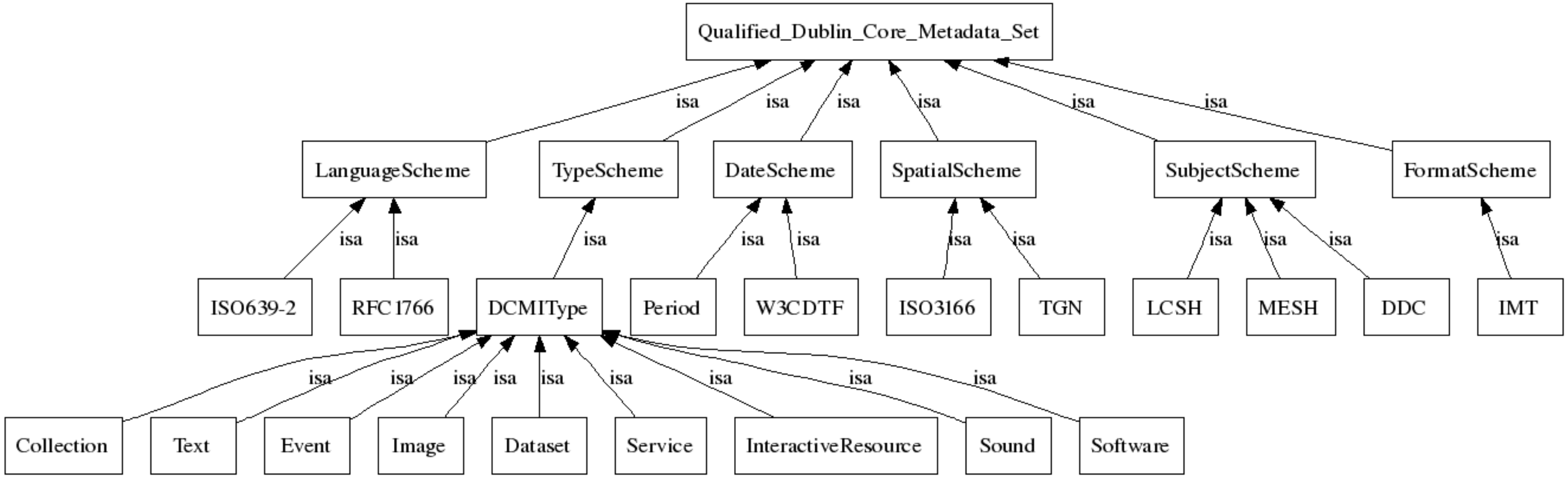


Figure 8: *Extensions of the Qualified Dublin Core.*

6.4 WordNet

WordNet is not a metadata schema, but an electronic database of word forms. It is widely used in linguistic research as well as information retrieval, and has been used in ontology research as well. It is perhaps the richest available source of information about terms that can be used as concepts in an ontology. Figure 9 shows a fragment of the different sorts of *picture* listed in WordNet. The image is taken from a web portal to WordNet available at `http://wordnet.princeton.edu/perl/webwn`.

The richness of WordNet makes it useful for complementing other metadata standards. For example, SUMO defines a set of mappings from WordNet, so that commonly used English words can (with some exceptions) be mapped to SUMO concepts. This would be useful in the Defence environment, where each user may have a different way of describing an image in natural language terms, and WordNet would provide a link between the different terms and some formal concept in the upper ontology.

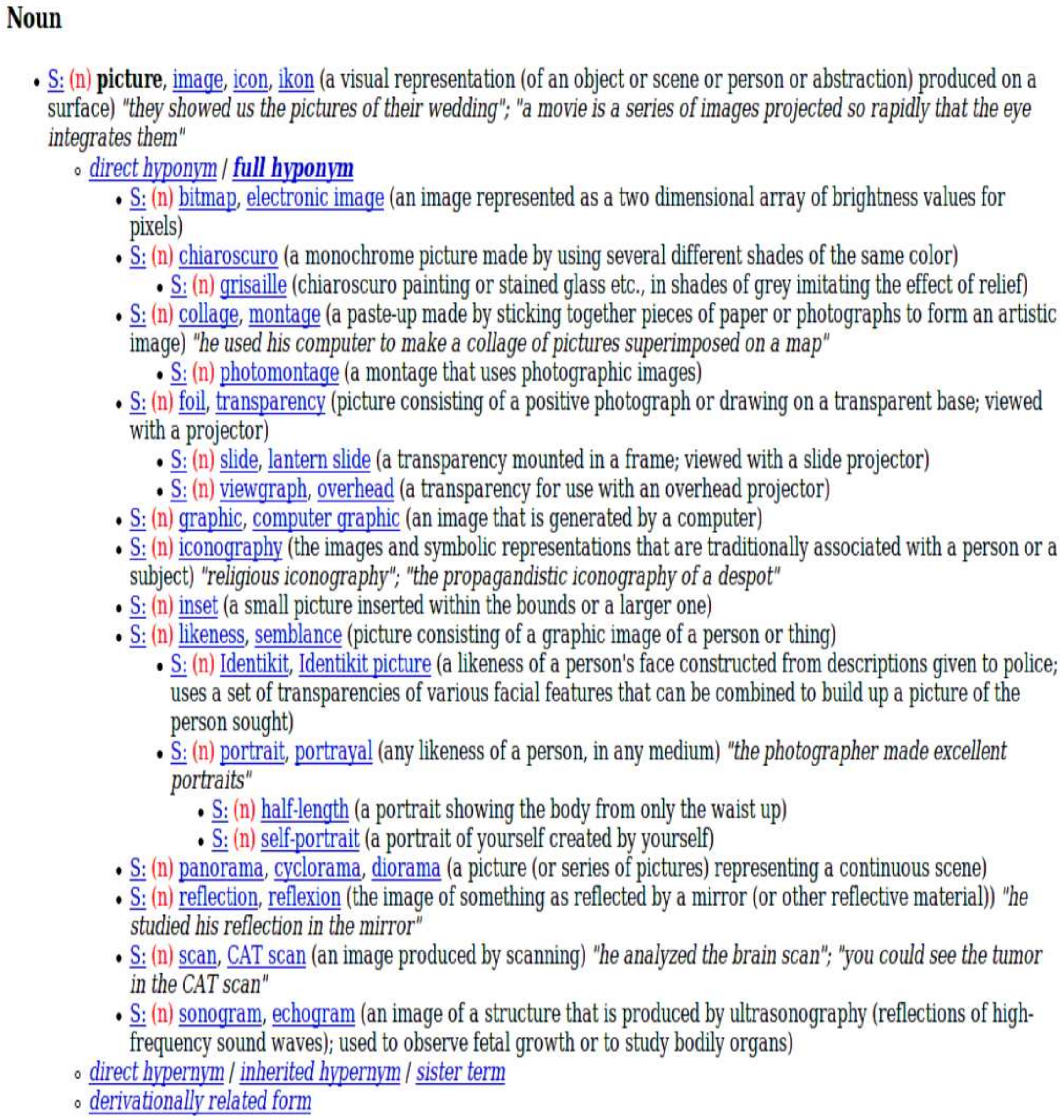
6.5 Image Data and W3C Best Practices

Image data will play a large part in the information disseminated in the AIR 7000 project. A large number of metadata types are available for image annotation, including VDO, MPEG-7 and EXIF. In addition, images lend themselves to a considerable amount of automatically generated metadata to do with the physical properties of the image, including its location, resolution, physical type, and so on. This kind of metadata can be distinguished from descriptive metadata about the content of the image, which is difficult to assign automatically.

Different users within the AIR 7000 user environment will have different requirements for image format and metadata practice, and the scope of this report doesn't cover detailing this information. Instead, it is worth considering the W3C best practices recommendations for image annotation.

The World Wide Web Consortium (W3C) has released a set of guidelines for annotating images on a large scale, using semantic technologies [van Ossenbruggen *et al.* 2006]. Admitting the difficulties inherent in the task, they flag a number of trade-offs that can affect the usefulness of annotations.

- *Production versus post-production annotation*
  "A general rule is that it is much easier to annotate earlier rather than later. Typically, most of the information that is needed for making the annotations is available during production time. Examples include time and date, lens settings and other EXIF metadata added to JPEG images by most digital cameras at the time a picture is taken, experimental data in scientific and medical images, information from scripts, story boards and edit decision lists in creative industry, etc. Indeed, maybe the single most best practice in image annotation is that in general, adding metadata during the production process is much cheaper and yields higher quality annotations than adding metadata in a later stage (such as by automatic analysis of the digital artifact or by manual post-production data)" [van Ossenbruggen *et al.*



Noun

- S: (n) **picture**, image, icon, ikon (a visual representation (of an object or scene or person or abstraction) produced on a surface) *"they showed us the pictures of their wedding"; "a movie is a series of images projected so rapidly that the eye integrates them"*
  - *direct hyponym* / ***full hyponym***
    - S: (n) bitmap, electronic image (an image represented as a two dimensional array of brightness values for pixels)
    - S: (n) chiaroscuro (a monochrome picture made by using several different shades of the same color)
      - S: (n) grisaille (chiaroscuro painting or stained glass etc., in shades of grey imitating the effect of relief)
    - S: (n) collage, montage (a paste-up made by sticking together pieces of paper or photographs to form an artistic image) *"he used his computer to make a collage of pictures superimposed on a map"*
      - S: (n) photomontage (a montage that uses photographic images)
    - S: (n) foil, transparency (picture consisting of a positive photograph or drawing on a transparent base; viewed with a projector)
      - S: (n) slide, lantern slide (a transparency mounted in a frame; viewed with a slide projector)
      - S: (n) viewgraph, overhead (a transparency for use with an overhead projector)
    - S: (n) graphic, computer graphic (an image that is generated by a computer)
    - S: (n) iconography (the images and symbolic representations that are traditionally associated with a person or a subject) *"religious iconography"; "the propagandistic iconography of a despot"*
    - S: (n) inset (a small picture inserted within the bounds or a larger one)
    - S: (n) likeness, semblance (picture consisting of a graphic image of a person or thing)
      - S: (n) Identikit, Identikit picture (a likeness of a person's face constructed from descriptions given to police; uses a set of transparencies of various facial features that can be combined to build up a picture of the person sought)
      - S: (n) portrait, portrayal (any likeness of a person, in any medium) *"the photographer made excellent portraits"*
        - S: (n) half-length (a portrait showing the body from only the waist up)
        - S: (n) self-portrait (a portrait of yourself created by yourself)
    - S: (n) panorama, cyclorama, diorama (a picture (or series of pictures) representing a continuous scene)
    - S: (n) reflection, reflexion (the image of something as reflected by a mirror (or other reflective material)) *"he studied his reflection in the mirror"*
    - S: (n) scan, CAT scan (an image produced by scanning) *"he analyzed the brain scan"; "you could see the tumor in the CAT scan"*
    - S: (n) sonogram, echogram (an image of a structure that is produced by ultrasonography (reflections of high-frequency sound waves); used to observe fetal growth or to study bodily organs)
  - *direct hypernym* / *inherited hypernym* / *sister term*
  - *derivationally related form*

Figure 9: *Part of the WordNet hierarchy for 'picture'.*

2006]). Obviously this implies that as much metadata as possible should be added during the initial posting in the information management process;

- *Generic vs task-specific annotation*
  A second problem involves the reason images might be annotated. If there is no specific task or application in mind, it is difficult to know what type of metadata should be used, in terms of content, abstraction, and so on. Generic annotation could be time consuming and costly, but even worse, might end up with metadata that is not suitable for specific applications down the line. A balance must be struck between application specific metadata and the ability to extend to future applications;

- *Different types of metadata*
  "While various classifications of metadata have been described in the literature, every annotator should at least be aware of the difference between annotations describing properties of the image itself, and those describing the subject matter of the image; that is, the properties of the objects, persons or concepts depicted by the image. In the first category, typical annotations provide information about title, creator, resolution, image format, image size, copyright, year of publication, etc. The second category describes what is depicted by the image, which can vary wildly with the type of image at hand. In many applications, it is also useful to distinguish between objective observations ('the person in the white shirt moves his arm from left to right') versus subjective interpretations ('the person seems to perform a martial arts exercise'). As a result, one sees a large variation in vocabularies used for this purpose" [van Ossenbruggen *et al.* 2006]. In addition, it is not uncommon that vocabularies might only define properties but leave possible values for those properties to be filled in with other vocabularies. As a result, more than one ontology is usually needed to annotate a single image, which is consistent with the approach we have taken throughout this document.

6.6 Summary

We have presented a number of existing schemas for the annotation of various kinds of data in the AIR 7000 operational environment. The relevance of each will need to be ascertained for the kinds of domain specific tasks that will be undertaken. The ADO_DM_MDP is an existing standard, and the first priority should be to translate it to RDF and investigate how it relates to SUMO concepts.

7 A Preliminary Architecture to Describe Metadata/Ontology Relationships

Figure 10 shows a proposed high level architecture for metadata annotation in AIR 7000, that allows various users to use different tools to add metadata to information objects. These users on the right hand side of the diagram (the 'workers with the shovels') would have discovered the information object based on the automatically assigned metadata such as location or time. They enrich the metadata set by adding domain and format specific metadata using some dedicated annotation tool. Meanwhile, the users on the left hand side have subscribed to all items described by particular metadata elements from the complete ontology. As these are added by the users on the right, they immediately become available to the users on the left through the subscription.

In this diagram, *Domain Ontology A* may correspond to the ontology used to formulate intelligence reports while *Domain Ontology B* might correspond to the ontology used by imagery analysts to describe features of images. The complete ontology is an aggregated or fused ontology that combines all the domain ontologies within the umbrella of an upper ontology.

This architecture would effectively provide for the merging of domain specific metadata into an umbrella scheme. When a user wants to retrieve data from the federated information environment, their retrieval activities are informed by the complete ontology rather than with one domain only, giving them access to the entire scope of relevant information.

Consider a very simple example to illustrate the functionality of the system. Suppose user A discovers an interesting image X based on its geographic position metadata. Using an image annotation tool (e.g. Photostuff) on the right hand side of figure 10, he annotates the image with the label 'tank'. This tag becomes immediately available in the repository, and allows user B (who subscribes to 'tank' or some superclass such as 'military vehicle') to discover the images. User B is more experienced in this domain and notices that this is a particular tank used in a known operation in the Gaza strip, and she annotates it as such. But at the same time analyst C has been annotating text with a tool such as Gate, and has tagged a different document Y with the same labels as user B tagged image X. Now the intelligence report in document Y can be immediately integrated with image X, to aid in developing the overall situation awareness.

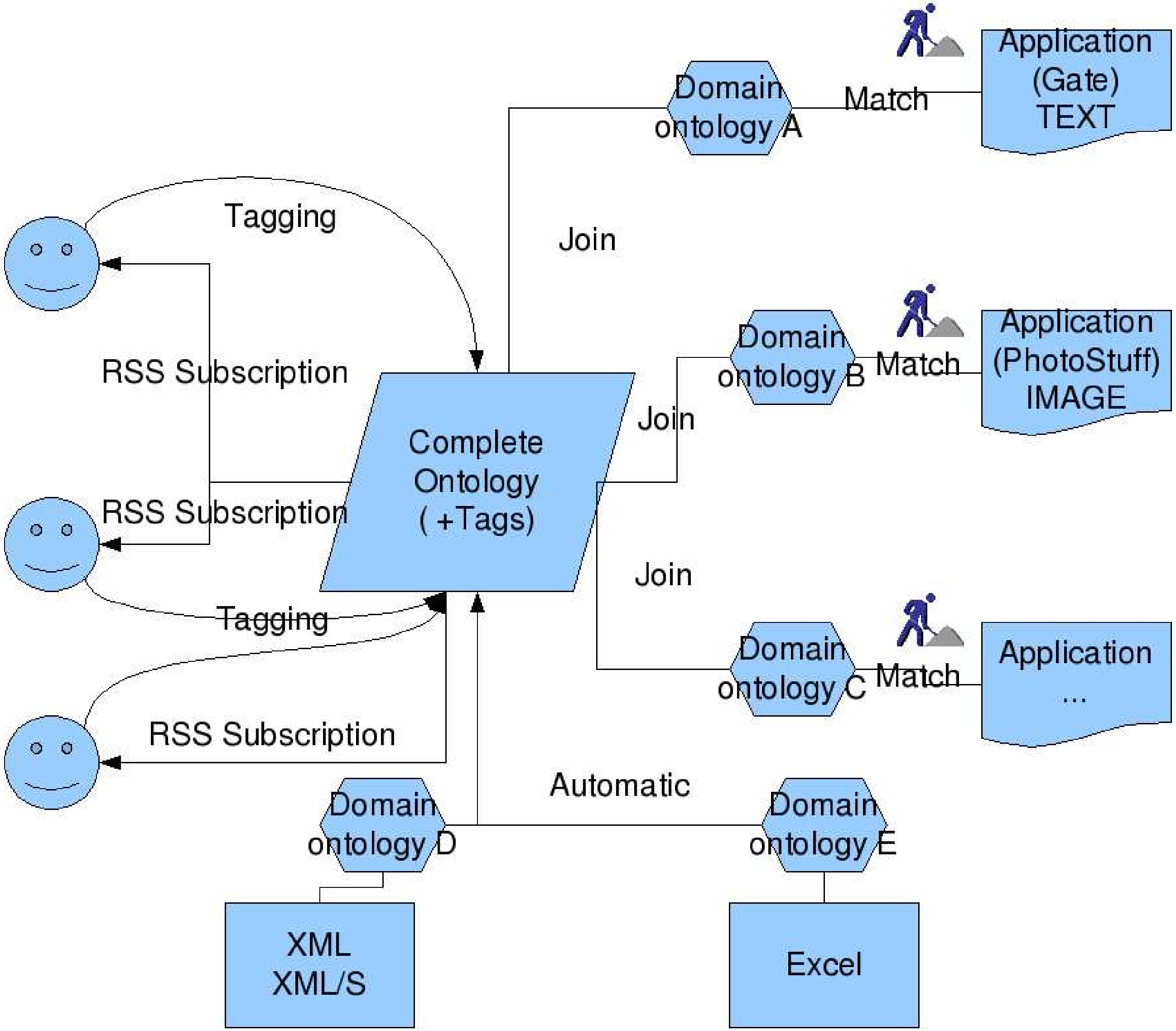


Figure 10: *A sketch of the possible architecture for metadata tagging and information dissemination.*

8 Summary and Recommendations

A large number of requirements for AIR 7000 information management either mandate or infer the use of metadata. Metadata provides a means for enabling data discovery and data management (including securing, organising, archiving and maintaining the integrity of data).

A large variety of metadata schemas currently exist, and each has its own utility within the scope of its intended use. However, interoperability between the schemas can be low. In the paper we discuss how the RDF data model together with ontologies provide the means for binding a number of extant schema together into a consistent framework. In addition, a discussion is made of the relationship between domain ontologies (such as DDMS and Dublin Core) and an upper ontology (such as SUMO or DOLCE). The latter helps to bring specific domain ontologies together into a consistent, unified framework of description. With an agreed upon upper ontology, existing ontologies will be able to be mapped to one another, allowing discovery across the full scope of information and not just within a particular user's community of interest. This is critical to support effective information pull from the federated information environment, but also to support proper collaboration and federated work practices.

In light of the discussion, the following recommendations are made:

- Work should begin immediately on identifying a set of relevant domain metadata for describing the various products (reports, plans, ISR data), processes and entities in AIR 700.
- AIR 7000 should commission a study to investigate how the various extant metadata standards should be made fully interoperable. In particular, any consideration of metadata and ontology must be done in a collaborative manner with each community of interest (covering planners, intelligence analysts, operators and so on) so that the final approaches are consistent and interoperable with wider Defence extant and future practices.
- RDF should be considered as the data model for storing metadata.
- SUMO should be considered as an upper ontology, with specific domain ontologies mapped onto SUMO.
- Future discovery services should adopt terms and processes consistent with the ontology adopted to support information management.
- An analysis of the practical risks associated with the use of metadata as a discovery enabler should be thoroughly examined. At this stage, limited work in this area has been undertaken.

A body of work in the application of ontologies and metadata to issues of data integration in military architectures exists. As part of the ongoing work we have already translated most of the DDMS to RDF, and are investigating its integration with SUMO. We focused initially on DDMS because it is comprehensive, yet relatively compact. We will tackle more complex standards such as the ADO_DM_MDP in future work. The immediate benefit will be to provide a capability to perform combined searches on documents marked up with these prominent standards in the United States and Australia. An extension of this work will provide the architectures recommended in this report for the AIR 7000 information management environment.

Acknowledgements

The authors are grateful for the feedback received from Thea Clark, Jeremy Manton, and Richard Bartholomeusz. We are particularly indebted to Simon Goss for his thorough and insightful comments as the referee.

References

. Chase G., Dall I. & Gani, R., Implications of the Global Information Grid for Australian network centric warfare, DSTO-TR-0697, 2006.

. Gilliland, A. J., Setting the Scene. Introduction to Metadata Pathways to Digital Information. Online Edition, Version 2.1 (`http://www.getty.edu/research/conducting\_research/standards/intrometadata/setting.html` last accessed April 29, 2008),

. Gilliland, T., Metadata and the World Wide Web. Introduction to Metadata Pathways to Digital Information. Online Edition, Version 2.1 (`http://www.getty.edu/research/conducting\_research/standards/intrometadata/metadata.html` last accessed April 29, 2008)

. Mack R., Ravin Y. & Byrd RJ., "Knowledge portals and the emerging digital knowledge workplace", IBM Systems Journal, vol. 40, no. 4, pp. 925 - 55, 2001.

. Manning, DP., Raghavan, P. & Schütze, H., Introduction to information retrieval, Cambridge University Press, Cambridge, 2007.

. Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., Schneider, L. WonderWeb Deliverable D17 The WonderWeb Library of Foundational Ontologies Preliminary Report ISTC-CNR c/o ISIB-CNR, C.so Stati Uniti, 4 35127 Padova Italy, 2007 (`http://www.loa-cnr.it/Papers/DOLCE2.1-FOL.pdf`)

. Ng, S., Dong, L., Gani, R., Smith, N. & Veres, C. Collection and Dissemination Architecture Study, Final Report, DSTO-CR-2007-0356, Defence Science and Technology Organisation, 2007

. Niles, I. and Pease, A. "Towards a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), Chris Welty and Barry Smith, eds., Ogunquit, Maine, October 17-19, 2001.

. Robertson, SE., "Computer retrieval", in BC Vickery (ed), Fifty years of information progress, Aslib, London, pp. 118 - 46, 1994.

. Semy, S. K., Pulvermacher, M. K., and Orbst, L. J. Toward the Use of an Upper Ontology for U.S. Government and U.S. Military Domains: An Evaluation. MITRE Technical Report, MTR 04B0000063, 2004

. Sowa, J. (`http://suo.ieee.org/email/msg13247.html` last accessed April 29, 2008)

. Stephens, RT., "Utilizing metadata as a knowledge communication tool", Proceedings of the International Professional Communication Conference, IEEE, Minneapolis, Minnesota, 2004.

. van Ossenbruggen, J., Troncy, R., Stamou, G. and Pan, J. Z. "Image Annotation on the Semantic Web: W3C Working Draft 22 March 2006" (`http://www.w3.org/TR/2006/WD-swbp-image-annotation-20060322/`)

. Varzi, A. C. Foreword to the special issue on temporal parts. The Monist, 83(3), 2000.

. Veres, C. “On the use of WordNet for semantic interoperability: towards ’cognitive computing’ ”. Proceedings of EMOI; CAiSE Workshops (3) 177-188, 2004.

DEFENCE SCIENCE AND TECHNOLOGY ORGANISATION DOCUMENT CONTROL DATA	1. CAVEAT/PRIVACY MARKING

2. TITLE An Approach to Information Management for AIR7000 with Metadata and Ontologies	3. SECURITY CLASSIFICATION Document (U) Title (U) Abstract (U)
4. AUTHORS Csaba Veres and Simon Ng	5. CORPORATE AUTHOR Defence Science and Technology Organisation Department of Defence, Canberra 2600, ACT

6a. DSTO NUMBER DSTO–TR–2289	6b. AR NUMBER	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE October, 2009

8. FILE NUMBER	9. TASK NUMBER AIR 7000	10. SPONSOR Director General Aerospace Development AIRCDRE John Oddie	11. No. OF PAGES 36	12. No. OF REFS 15

13. URL OF ELECTRONIC VERSION http://www.dsto.defence.gov.au/corporate/reports/DSTO–TR–2289.pdf	14. RELEASE AUTHORITY Chief, Joint Operations Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for Public Release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SOUTH AUSTRALIA 5111

16. DELIBERATE ANNOUNCEMENT

No Limitations

17. CITATION IN OTHER DOCUMENTS

No Limitations

18. DSTO RESEARCH LIBRARY THESAURUS

AIR7000	metadata
rdf	ontology
information management	

19. ABSTRACT

This paper discusses the concept 'metadata', and shows its importance in information collection and dissemination activities. We also show that the information management components of maritime patrol and response mandate the effective use of metadata. We then propose an approach based on Semantic Technologies including the Resource Description Framework (RDF) and Upper Ontologies, for the implementation of metadata based dissemination services for AIR 7000. A preliminary architecture is proposed. While the architecture is not yet operational, it highlights the challenges that need to be overcome in any solution to the information management tasks of AIR 7000, and provides a possible form for the solution.